IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 21, 2019, accepted December 12, 2019, date of publication December 30, 2019,
date of current version January 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962154

A Natural Language Process-Based Framework for Automatic Association Word Extraction

ZHENG HU[1], JIAO LUO[1], CHUNHONG ZHANG[2], AND WEI LI[3]
[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3]Laboratory for Intelligent Networks and Systems, Northern Illinois University, DeKalb, IL 60115, USA

Corresponding author: Zheng Hu (huzheng@bupt.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFF0302601, and in part by the Creative Research Groups of NSFC under Grant 61421061.

ABSTRACT Word association, revealing mental representations and connections of human, has been widely studied in psychology. However, the scale of available associative cue-response words is severely restricted due to the traditional manually collecting methodology. Meanwhile, with the tremendous success in Natural Language Process (NLP) tasks, an extremely large amount of plain texts can be easily acquired. This suggests an insight about the potential to find association words automatically from the text corpus instead of manually collection. As an original attempt, this paper takes a small step toward proposing a deep learning based framework for automatic association word extraction. The framework mainly consists of two stages of *association word detection* and *machine association network construction*. In particular, attention mechanism based Reading Comprehension (RC) algorithm is explored to find valuable association words automatically. To validate the value of the extracted association words, the correlation coefficient between semantic similarities of machine and human association words is introduced as an effective measurement for evaluating association consistence. The experiments are conducted on two text datasets from which together about 20k association words, more than the existing largest human association word dataset, are finally derived. The experiment further verifies that the machine association words are generally consistent with human association words with respect to semantic similarity, which highlights the promising utilization of the machine association words in the future researches of both psychology and NLP.

INDEX TERMS Word association, natural language process, semantic similarity, attention mechanism.

I. INTRODUCTION

What is the first responding word coming into mind when one person is given the word *coffee*? This is an interesting mental capability that has long been studied under the terminology *Association* thinking. In psychology, association typically refers to a mental connection among different appearances due to some inducements [1], which can be prominently perceived and revealed through the phenomenon of *word association*. As a typical example, *Brazil* is usually observed as *response* word to the *cue* word *coffee*, which would be persuasively explained by substantial common knowledge that the Brazil is generally viewed as one of the famous countries of global coffee production.

While the **cue-response** word pattern has long been widely utilized for research works beyond psychological speculation, unfortunately, it is usually quite expensive and time consuming to collect association words. For example, the project [2] manually collected an English Word Association dataset between year 2011 and 2018, consisting of $+12,000$ cue and $3,684,699$ responses words from over $90,000$ subject participants recruited online. The methodology of automatic association word collection without human effort becomes more critical in artificial intelligence applications.

Moreover, the majority of the published Word Association datasets are collected under controlled experimental environments. Such cue-response words might be significantly different to what people associate in real scenarios. When one scans a news website, for instance, the particular news content and sentimental bias would allow alternative cue-response words not consistent with those collected from experiments. Surprisingly, capturing cue-response words explicitly when human response words upon a given text other than a single cue word has not been well studied.

The associate editor coordinating the review of this manuscript and approving it for publication was Soon Xin Ng.

Meanwhile, an extremely large amount of text corpora are easily available online nowadays, suggesting promising potential to explore association words directly from daily documents. One sensible kind of text is news article and its comments, which consist of diverse review responses from various audiences. The association is commonly observed in that the sentences of comment tend to contain words not directly mentioned but semantically related to the news article. In other words, the news article and comments could be viewed as reservoir for automatic cue-response word collection. Considering a sentence "*Smartphone brand OnePlus launched the fifth generation of the main mobile phone today*" in a news article and one of its comments "*The Infinity Display of Samsung is amazing*". Three cue-response words, *"smartphone-Samsung", "smartphone-Infinity Display" and "OnePLus-Samsung"* can be recognized from them. That is, the cue word "smartphone" can easily remind people of company "Samsung" as well as the smartphone function feature "Infinity Display". Also, one smartphone brand "OnePlus" could stimulate association to another brand "Samsung".

Automatic detection of precise and useful cue-response words from text is difficult. There has been extensive studies about inherent word relationship in the literature of NLP. One primary strategy of these approaches is to take word co-occurrence statistics as measurement of latent semantic relationship among words [3]–[5]. However, these approaches are less suitable for association word detection. One reason is that the simple lexical co-occurrence reflects many other factors besides association. Another problem is that given a particular cue word in article, it is hard to decide which response words might be related to it since the comment is stimulated by the whole article but not any single word in it. Therefore, the nature of many-to-many correspondence between article words and comment words raises a major challenge for meaningful association word finding.

To the best of our knowledge, there is no study focusing on automatic association word detection from plain text. This paper seeks to lay foundation on which this problem is solved. First, a framework of finding association words by introducing Reading Comprehension (RC) algorithm of NLP is proposed. Then, to analyze the associative properties of the found words, a *machine association network* is constructed, where each node is a word and each edge connects two words of cue-response pair. Two association networks are obtained with roughly 100k nodes from two text datasets respectively. To distinguish, the network constructed upon association words manually collected from participants is denoted as *human association network*. Finally, the semantic similarity of machine and human association networks is estimated to demonstrate their consistence from the perspective of association.

In particular, the critical issue of detection of one-to-one cue-response words from many-to-many words of text is framed as an attention model between two sequential text encoders. The *attention mechanism* is adopted to capture how a cue word is responsible for response word. The idea behind is that the notions of *attention mechanism* and *human rationale* are closely related and both of them highlight the word importance [6]. Pappas and Popescu-Belis [7] demonstrate that there is a positive correlation between human-annotated attention and attention weights induced by the models. This suggests that the attention mechanism can be utilized as proxy to guide the detection of association words, under the assumption that human association is rational. As a result, the learned *attention weight* is utilized to characterize the association strength between cue-response words.

The contributions of this paper are summarized as follows:(1) Proposing a framework of automatic association word detection from plain text based on two neural network sequential encoders with attention mechanism. (2) Constructing machine association network using attention weight learned from cue-response words of two text datasets. (3) Verifying the consistence between machine association network and human association network through semantic similarity, which indicates promising value of machine association network for many research fields.

The reminder of the paper is organized as follows. Section II surveys related works in word association, semantic similarity and attention interpretability. Section III describes the extracting process of association words by achieving NLP tasks and the construction process of machine association network. Section IV introduces three measurements to estimate semantic similarity between cue-response words based on machine association network. Section V discusses the experimental results and section VI draws conclusions.

II. RELATED WORKS

A. WORD ASSOCIATION

There have been many previous works studying on the association in the traditional psychological way. Mednick developes the Remote Associates Test [8], [9] requiring participants to find a mediating word that can link the given three distinct test words. Word-Association tests were brought up and used in [10]. This test usually asks for the first associated word that comes to mind when given a set of words. Another novel work [11] contrasts the word similarity derived from both human word association network and word embedding model, and finds out the results of former are more close to the ground truth judgements made by human. Recently, a large scale human association dataset named SWOW is collected manually for over 12000 cue words in [2]. The utility of the dataset in several different contexts, including lexical decision and semantic categorization is also evaluated. In this paper, a framework automatically extracting association words from plain texts is proposed to facilitate theoretical study of psychology.

B. SEMANTIC SIMILARITY

This paper tries to understand the semantic property of the machine association network. Semantic similarity is a central

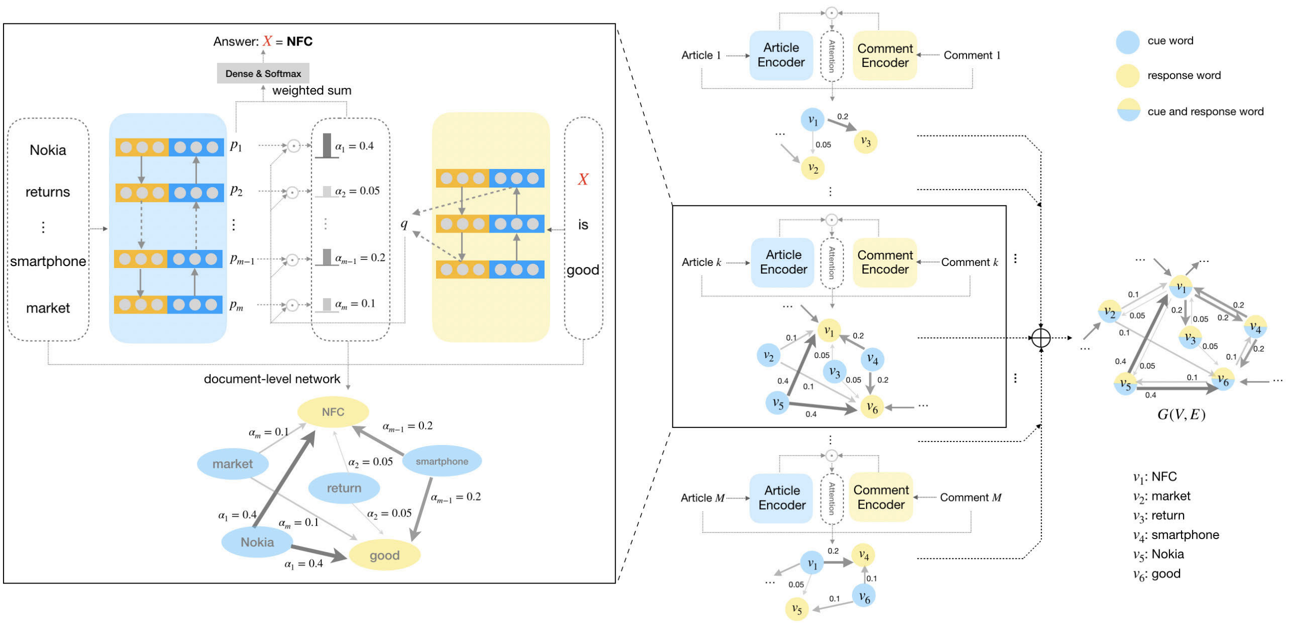

FIGURE 1. The framework of automatic association word extraction from plain text based on two sequential encoders and an attention model. The blue node and yellow node represent the cue word and response word respectively. The node filled with half blue and half yellow indicates this node word is a cue word as well as a response word. The thickness of the directed edge from cue word to response word indicates their association strength. The number marked near the edge represents the association strength between the two node words connected by the edge.

concept in many cognitive theories of language. In a series of studies, fMRI evidence has shown that the distributed lexical semantic model can predict the activation patterns of different brain regions when reading common words [12]. Researches in semantic similarity modeling human associations tend to sort into three poles. The first one is distributional similarity-based methods such as Explicit Semantic Analysis (ESA) [13] and Salient Semantic Analysis (SSA) [14] representing a word by the surrounding context it keeps. There are also metrics based on large corpora such as Pointwise Mutual Information (PMI) and second order PMI [15]. The third kind relies on resources such as thesauri or lexicon [16]. De Deyne *et al.* [11] propose a spreading activation approach to predict semantic similarity of human word association network. In this paper, the approach described in [11] is adopted as the random walk measure to model semantic similarity over the word association network of the machine.

C. INTERPRETABILITY OF ATTENTION

The proposed framework extracts association words automatically from plain text mainly based on an attention model. There have been attempts to provide insights into the interpretability of attention. The proposed first-derivative saliency of attention succeeds in pinpointing how the alignment process contributes to the final predictions [17]. Alvarez-Melis and Jaakkola [18] provide explanations of black-box predictions via a causal framework. Several pieces of work in natural language inference try to understand their models by simply visualizing the attention layer [19]. Unfortunately the similarities between hidden states represented by a heat-map often reveal little information towards the decision. Another line of work tends to involve human rationales in the process of understanding the decisions of attention-based models. Pappas and Popescu-Belis [7] demonstrate that there is a positive correlation between human-annotated attention and induced attention weights. Building a mapping from human rationales to attention weights to guide models is another promising avenue [6]. The focus of this paper is to utilize attention mechanism as a tool to acquire the internal association relationship for cue-response word, instead of interpreting how the inputs to which the model assigned large attention weights are responsible for outputs.

III. MACHINE ASSOCIATION NETWORK

The focus of this study is to automatically extract association words from plain text and then construct the corresponding machine association network. The success of NLP inspires the consideration that the semantic interaction modelling capability of NLP could eventually enable the acquirement of association patter among words. The extracting framework is formulated based on the RC task as shown in Fig. 1. The input of the framework is a text corpus consisting of article-comment pair, each one of which could be used as a source of meaningful cue-response candidate words. The output is a machine association network consisting of the resulting words with corresponding association strength.

To extract cue-response words, two neural network based sequential encoders are adopted to model the contextual and semantic information of article and comment respectively (III-A). The attention mechanism is utilized to capture how and to what extent an article word is associated to a comment word (III-B). A document-level association network is constructed with the association words and their attention weights extracted from one article-comment pair (III-C.1). Accordingly, a corpus-level association network is then obtained by integrating all the document-level networks (III-C.2).

A. ASSOCIATION DETECTION

The extracting process of association words is inferred by achieving NLP task on article-comment corpus since the model of NLP task is capable of detecting and understanding the underlying linguistic relationship between words. Here the natural language reading comprehension task is employed to create associative mapping from cue word to response word. Under the guidance of RC task, the learned interaction between words by attention mechanism is consistent with human associative rationale to a certain extent [7].

The RC task seeks to estimate the conditional probability $\Pr(a|p, q)$, where p is an article, q a query relating to that article, and a the correct answer to that query [20], which is expected to have higher probability than all other words. When used on the article-comment text, the sentence of comment is viewed as query q with one entity replaced with a placeholder, and the answer a is the replaced entity. For instance in Fig. 1, the comment sentence "*NFC is good*" is rephrased as "*X is good*" by replacing the entity "NFC" with a placeholder X. It is believed that the replaced word "NFC" would be predicted as correct answer once the RC task can read and understand the article's sentence "*Nokia return ... smartphone market*". By predicting the entity "NFC", the cue-response words "Nokia-NFC" and "smartphone-NFC" can be detected by the attention weights learned by the RC model.

To be specific, the classic and simple *Attentive Reader* RC model [21]–[23] is utilized to detect the association words. As shown in the left part of Fig. 1, given the (article, comment, answer) triple (p, q, a), $p = \{p_1, \ldots, p_m\}$ and $q = \{q_1, \ldots, X, \ldots, q_l\}$ are sequences of tokens for article and comment sentences, with q containing only one placeholder token X replacing answer a. The goal is to refer the replaced token $a \in p$ as correct answer. The association detection model consists of two sequential encoders and an attention block, which are described in the following three steps:

1) TEXT ENCODING

The article p is encoded by a bi-directional Long short-term memory (LSTM) network to learn a hidden vector for each word, presenting its contextual feature in the article. The outputs of the forward and backward LSTMs for word at position i are denoted as $\overrightarrow{\boldsymbol{h}}_i \in \mathbb{R}^h$ and $\overleftarrow{\boldsymbol{h}}_i \in \mathbb{R}^h$ respectively. The composite contextual representation $\tilde{\boldsymbol{p}}_i \in \mathbb{R}^{2h}$ for each word p_i contains the right and left contextual information by concatenating $\overrightarrow{\boldsymbol{h}}_i$ and $\overleftarrow{\boldsymbol{h}}_i$,

$$\overrightarrow{\boldsymbol{h}}_i = \mathrm{LSTM}(\overrightarrow{\boldsymbol{h}}_{i-1}, p_i), \quad i = 1, \ldots, m \tag{1}$$

$$\overleftarrow{\boldsymbol{h}}_i = \mathrm{LSTM}(\overleftarrow{\boldsymbol{h}}_{i+1}, p_i), \quad i = m, \ldots, 1 \tag{2}$$

$$\tilde{\boldsymbol{p}}_i = \overrightarrow{\boldsymbol{h}}_i || \overleftarrow{\boldsymbol{h}}_i \tag{3}$$

Similarly, the embedding $\tilde{\boldsymbol{q}} \in \mathbb{R}^{2h}$ of a comment q with length l is formed by the concatenation of the final forward and backward outputs of another bi-directional LSTM,

$$\tilde{\boldsymbol{q}} = \overrightarrow{\boldsymbol{h}_l} || \overleftarrow{\boldsymbol{h}}_1 \tag{4}$$

Note that the $\tilde{\boldsymbol{q}}$ is the embedding of the whole comment q instead of individual word of it, as that doing for each article word p_i. The reason will become apparent as the attention mechanism is described.

2) ASSOCIATION STRENGTH ESTIMATION

So far the article-comment are encoded into vectors. Next, the attention mechanism is utilized to compute the associative interaction between each word $\tilde{\boldsymbol{p}}_i$ of article and comment $\tilde{\boldsymbol{q}}$ by selecting article word closely related to the comment, resulting in an attention weight α_i.

$$\alpha_i = f_{\text{attention}}(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}) \tag{5}$$

where many variants for the abstract attention estimation function $f_{\text{attention}}$ will be applied to test its core capability of association detection. As attention mechanisms pay more attention to these article words that are more associated with the comment, the weight can be interpreted as the strength to which the model associates a particular word p_i in article to the underlying process logic when generating the comment q. More intuitively speaking, when given a comment, the model is more likely to associate with the article words receiving larger weights, while less likely to associate with those words assigned with smaller weights. Of key importance is that this attention weight α_i is the basis for the construction of machine association network later.

Recall that the attention weight α_i is between a passage word p_i and the whole comment sentence q but not any individual comment word. Obviously, this can not be directly used to detect cue-response words. To tackle this problem, the attention weight α_i is decomposed from a word-to-sentence level into a word-to-word level. A simple decomposing principle is adopted that each comment word q_j is assumed having equal attention weight to one article word p_i. That is, the attention weight between words p_i and q_j is $\alpha_{ij} = \alpha_i$. It is noted that although there is RC model named *Impatient Reader* [20] that can estimate the attention weight directly on word-to-word level. The Impatient Reader is not adopted in this paper since Impatient Reader and Attentive Reader (the adopted one) perform comparably in the RC task but the former is more computationally complex.

3) ANSWER PREDICTION

Finally to predict the answer, all the word embeddings $\{\tilde{\boldsymbol{p}}_i\}$ are combined into an output vector $\boldsymbol{o}$ upon attention weights to represent the holistic context of the article, i.e., $\boldsymbol{o} \in \mathbb{R}^{2h} = \sum_i \alpha_i \tilde{\boldsymbol{p}}_i$, which is then used to predict the answer via a fully-connected layer with softmax function,

$$\boldsymbol{s} = \text{softmax}(\boldsymbol{W}_a^\top \boldsymbol{o}) \tag{6}$$

where vector $\boldsymbol{s}$ defines a probability distribution over the distinct entities which appear in the passage p and $\boldsymbol{W}_a \in \mathbb{R}^{2h \times d}$ is the network parameter, each element of which is randomly initialized and learned through the training of a negative log-likelihood objective.

The probability of a particular candidate entity $e_i \in p$ as being the correct answer is proportional to one element of vector **s**,

$$\Pr(e_i|p, q) \propto s_i \tag{7}$$

Finally, the candidate entity with the maximum probability is selected as the predicted answer,

$$a^* = \underset{e \in p}{\text{argmax}} \ \Pr(e|p, q) \tag{8}$$

Once the correct answer a^* is predicted, the resulting attention weights learned by the model would be appropriate to detect word association.

B. ATTENTION MECHANISM

As aforementioned, the choice of attention algorithm is critic to the result of association detection. Meanwhile, it is observed that the *human* association dataset generally appears stability on certain association patterns to some extent. For instance, it is always found that the *"coffee-Brazil"* word pair is more closely related than *"coffee-rocket"*. Therefore, the *machine* association words are expected to also possess similar stability without remarkably varying with respect to different attention function. That is, the attention mechanism is assumed capable of establishing relatively stable and reasonable associative patterns regardless of which particular attention algorithm is adopted. This hypothesis is verified in later experiment by employing three different attention algorithms as $f_{\text{attention}}$ function in (5). Below the detail of the three attention alternatives are briefly described.

1) DOT-PRODUCT ATTENTION

Dot-Product attention is one of the most commonly used attention algorithms. It is originally used in machine translation by Luong *et al.* [24]. The attention function $f_{\text{attention}}$ is computed by a simple dot product of the word vector $\tilde{\boldsymbol{p}}_i \in \mathbb{R}^{2h}$ and the comment vector $\tilde{\boldsymbol{q}} \in \mathbb{R}^{2h}$:

$$f_{\text{attention}}(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}) = \text{softmax}_\text{i}(\tilde{\boldsymbol{p}}_i^\top \tilde{\boldsymbol{q}}) \tag{9}$$

2) BILINEAR ATTENTION

Bilinear attention [24] is another commonly used attention algorithm. Chen *et al.* [21] find it effective when applied in RC. This mechanism is achieved by multiplying a trainable parameter matrix $\boldsymbol{W}_s \in \mathbb{R}^{2h \times 2h}$ between the dot product of $\tilde{\boldsymbol{p}}_i$ and $\tilde{\boldsymbol{q}}$:

$$f_{\text{attention}}(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}) = \text{softmax}_\text{i}(\tilde{\boldsymbol{p}}_i{}^\top \boldsymbol{W}_s \tilde{\boldsymbol{q}}) \tag{10}$$

3) MLP ATTENTION

Multi-layer perceptron (MLP) attention [25] is an additive attention rather than a dot-product one. It is used to improve the performance of machine translation in [25]. In contrast to the bilinear attention, this mechanism makes use of a tanh layer on the top of a multi-layer perceptron instead of a bilinear term to compute $f_{\text{attention}}$ as

$$\begin{aligned} \boldsymbol{m}_i &= \tanh(\boldsymbol{W}_p \tilde{\boldsymbol{p}}_i + \boldsymbol{W}_q \tilde{\boldsymbol{q}}) \\ f_{\text{attention}}(\tilde{\boldsymbol{p}}_i, \tilde{\boldsymbol{q}}) &= \text{softmax}(\boldsymbol{W}_m^\top \boldsymbol{m}_i) \end{aligned} \tag{11}$$

where $\boldsymbol{W}_p \in \mathbb{R}^{2h \times 2h}$, $\boldsymbol{W}_q \in \mathbb{R}^{2h \times 2h}$ and $\boldsymbol{W}_m \in \mathbb{R}^{2h}$.

C. MACHINE ASSOCIATION NETWORK

Once acquiring the attention weight between article word p_i and comment word q_i, the machine association network can be constructed on two hierarchical levels: *document-level* and *corpus-level*. The document-level network is a local directed graph where the nodes consist of cue-response words extracted from a single article and all its comments, while the corpus-level network is a global directed graph integrated over all the document-level networks. The corpus-level network is utilized as a feasible source to test the similarity properties of machine association words.

1) DOCUMENT-LEVEL ASSOCIATION NETWORK CONSTRUCTION

In general, the way of constructing machine association network follows the principle of human association network [2], [26]. Denote here the cue-response word pair extracted from an individual article d^k as $(p_i, q_j)_k$ for cue word p_i and response word q_j with attention weight α_{ij}^k, where k stands for the index of an article. First, if either article word p_i or comment word q_j is stop word or punctuation, the $(p_i, q_j)_k$ is removed such that the retained association words are meaningful. Notice that the article d^k may contain more than one association pair $(p_i, q_j)_k$ with the same cue-response vocabulary words. In this case, they are integrated into only one $(p_i, q_j)_k$ with their α_{ij}^k added together to represent the *association strength* f_{ij}^k of the association pair $(p_i, q_j)_k$. That is,

$$f_{ij}^k = \sum_{\substack{s|p_s=p_i \\ t|q_t=q_j}} \alpha_{st}^k \tag{12}$$

where the association strength f_{ij}^k is an indicator of how ease the response word is stimulated by the cue word, to some extent the same role of *association frequency* for human association words. Taking an example to illustrate the process more clearly, the cue word p_i *Trump* may present three times in the article d^k and the response word q_i *president* may appear once in the corresponding comment. Noting that there is an attention weight between every cue-response pair, it means that there are three distinct attention weights, assumed as α_1, α_2 and α_3, between *Trump* and *president* in the article d^k. Therefore, at the level of a single document d^k, the association strength for the association pair $(Trump, president)_k$ is $f^k_{(Tru,pre)} = \sum_{i=1}^{3} \alpha_i$.

A document-level association network $G_k = (V_k, E_k)$ can be constructed based on the set of $((p_i, q_j)_k, f^k_{ij})$. Each node $v \in V_k$ represents a word of either p_i or q_j, a directed edge $e \in E_k$ is created from node v_i to node v_j with edge weight f^k_{ij} if word q_j is a response to word p_i as a cue in article k. Normally, the scale of the constructed network G_k might be large if all the original cue-response pairs are contained by G_k, especially when the article k is long or has many comments. In a consideration to make the G_k less bulky, for each cue word p_i, only the top 30 response words with the highest association strength f^k_{ij} are kept. It is worth noting that for the establishment of the human association network, in contrast, only the first 3 response words that are most easily associated are collected [2]. Thus, it is believed that choosing the top 30 response words with the highest association strengths is sufficient and can effectively capture the association properties conveyed by the text. Besides, keeping only the largest strongly associated words in G_k is high likely to ensure that the retained nodes both have ingoing and outgoing edges. That is, the network G_k is connected graph and there is always a path between arbitrary vertex pair, which will be convenient for the analysis method of association semantic property based on graph.

2) CORPUS-LEVEL ASSOCIATION NETWORK CONSTRUCTION

However, the G_k built on a single article and its comments is rather a local and limited association evidence. In this network, most nodes from comments are identical to each other in terms of association behavior since they share almost the same neighboring article nodes. It is sensible to construct a corpus-level machine association network based on all the document-level networks. To achieve this, all the local network $\{G_k\}$ of the corpus with M articles are integrated into a global network $G' = (V', E')$. The V' is the union of node sets $\{V_k\}$ and E' is the union of edge sets $\{E_k\}$ from all G_k. The edge weight f_{ij} of $e_{ij} \in E'$ is simply the accumulation of all the edge weights $\{f^k_{ij}\}$, that is,

$$V' = \cup V_k, E' = \cup E_k, \quad k = 1, \ldots, M$$
$$f_{ij} = \sum_{k=1}^{M} f^k_{ij}, \forall e_{ij} \in E' \tag{13}$$

where the f_{ij} reflects the global association strength between cue-response $(p_i, q_j)_k$ over the entire corpus.

According to the principle of human association network construction, the self-loops caused by the identical cue and response words are removed from $G'(V', E')$. Moreover, to ensure that all words can be reached by both in-going and out-going edges, only response nodes that also occur as cue nodes are retained. This means the nodes with zero out-degree are removed, and the largest strongly connected component is extracted by keeping cue nodes that appear at least once as response nodes. The final result of these constructing operations leads to a directed and weighted graph $G(V, E)$, namely the *machine association network*, from the intermediate graph $G'(V', E')$.

IV. ASSOCIATION SEMANTIC PROPERTY ANALYSIS

Many researches have demonstrated that semantic similarity plays an important role in characterizing human association network [2], [11], [27]. Hutchison *et al.* [27] find that in priming, human brain's latency to process a response word could be predicted by the semantic similarity between the cue word and the response word. Another line of work suggests that when people evaluate word similarities, their results are more close to the semantic similarity estimated over the word association network than that from word embedding models [11]. Recently, sufficient experiments are conducted to show the high correlation between human judged similarity and similarity derived from word association network [2]. Against this backdrop, it seems not only sensible but also necessary to understand the semantic similarity property of the machine association network $G(V, E)$, to appropriately assess its potential utility on the study of psychology as that of human association network.

Here, three measurements to estimate semantic similarity between cue-response words based on machine association network $G(V, E)$ are explored. These three measures, *Associative Strength* (AS), *Positive Pointwise Mutual Information* (PPMI), and a *Random Walk Similarity* (RWS) measure, are different from each other in terms of amount of information they use. The simplest and most common approach is to consider only the direct neighboring nodes between two words. In this case, two words will obtain a more similar meaning if they share more direct neighboring nodes. However, extending the concept of similarity to include indirect paths connecting two words to capture a more comprehensive and global measure of similarity is quite intuitive and straightforward. As a result, both scenarios will be covered in the following sections.

In general, in each of the three measurements, each word gets a vector representation $\mathcal{F}$. The semantic similarity of two words w_i and w_j is defined as the cosine similarity of the their vector representations,

$$\text{Sim}(w_i, w_j) := \frac{\mathcal{F}_i \cdot \mathcal{F}_j}{\|\mathcal{F}_i\| \times \|\mathcal{F}_j\|} \tag{14}$$

3) ASSOCIATIVE STRENGTH (AS)

The roughest way of measuring semantic similarity of two words w_i and w_j is to estimate the common neighbours they sharing in the machine association network $G(V, E)$. To implement this, L1-norm normalization is performed over all the outgoing edge weights of each node, i.e.,

$$\tilde{f_{ik}} = \frac{f_{ik}}{\sum_{k'} f_{ik'}} \tag{15}$$

In the AS measurement, the vector representation $\mathcal{F}_i$ of word w_i with length $|V|$ is then created where the element at position k is $\tilde{f_{ik}}$. The associative strength similarity AS is then calculated as defined in (14).

4) POSITIVE POINTWISE MUTUAL INFORMATION (PPMI)

Recchia and Jones have shown that the PPMI can be used to predict the behavior in various language processing tasks [28]. As an information theoretic measure based on the full distribution of the response to a cue word, PPMI is suggested to be reasonably successful as an enhanced measure of associative strength. This paper follows the definition of PPMI measures in [2], which is given by

$$\mathrm{PPMI}(w_k|w_i) = \max(0,\ \log_2(\frac{\tilde{f_{ik}}}{\sum_i' \tilde{f_{i'k}}})) \tag{16}$$

In the PPMI measurement, $\mathrm{PPMI}(w_k|w_i)$ is the k_{th} element of the vector representation $\mathcal{F}_i$ of word w_i. Again, the semantic similarity of two words is calculated as the inner product between their vector representations, just as (14) does. Compared with the similarity AS considering only the direct neighbors, PPMI takes the edge weight distributional information derived from the entire association network into account. In this case, responses nodes which are frequently given for many cue nodes are thought to be less informative than responses given only for a small number of cues.

5) RANDOM WALK SIMILARITY (RWS)

From a more *global* point of view, it is found that the semantic similarity of two word nodes could be also reflected by the indirect paths through which they are connected, in addition to the direct connections used by the above two similarity measures. A random walk process similar to the Katz index [29] is taken for this principle. Two words are considered to be more similar in the case that there are more short paths connecting them, because it is easier to start at one word node and end at the other from a random walk through the graph. Given the maximum length r of random walk, the produced transition matrix $\boldsymbol{P}_r$ is computed as

$$\boldsymbol{P}_r = \sum_{i=0}^{r-1} (\alpha \boldsymbol{P}_G)^i \tag{17}$$

where $\alpha (= 0.75)$ is a damping parameter [30] and $\boldsymbol{P}_G$ is the transition matrix of the machine association network G. The element $p_{r_{ik}}$ of the produced $\boldsymbol{P}_r$ is regarded as the k_{th} element of the vector representation $\mathcal{F}_i$ of word w_i in the RWS measurement. Finally, in light of (14), the inner product of $\mathcal{F}_i$ and $\mathcal{F}_j$, is considered as the similarity of two words.

V. EXPERIMENTS AND INSIGHTS

As aforementioned, the value of extracted machine association network is mainly evaluated in terms of whether it is inherently consistent with human association network on semantic property. To investigate the potential consistence, the semantic similarities derived from both machine association network and human association network are quantitatively analyzed as validation metrics. In particular, the correlation coefficient is used to measure the statistic relationship between the two semantic similarities of machine and human association networks. The effect of the attention mechanism is also evaluated to see whether diverse ways in which machine performs association will influence the stability of semantic property of machine association network. In addition, to understand the association behavior of models that *grow* in different corpus environments, experiments are carried out on two datasets with different inherent logical relatedness between article and comment.

A. DATASETS

1) MACHINE ASSOCIATION DATASETS

Recall that there is no dataset available originally for the task of association word extraction. Thus, the experimental datasets are borrowed from the NLP domain of RC and some revisions are conducted on these datasets to meet the requirement of association detection. The first dataset is CNN [20], which is a typical RC dataset consisting of 387, 420 passage-query-answer samples. To align the terminology with the association detection task described in section III-A, the passage-query-answer sample is viewed as article-comment-answer form. That is, the framework will create appropriate association relationships between words from passage and query through finding the correct answer.

The second dataset is NYT, consisting of news articles and all their corresponding comments collected from *New York Times*.[1] To make the original comment sentence satisfy the required form of model training, a comment entity(word) which also appears in its corresponding article body is randomly selected and is replaced with a placeholder. In other words, the selected entity is the missing entity namely the answer model needs to infer. In this way, a comment is converted into a query. Since there is usually more than one comment for an individual news article, for each comment satisfying the condition that its entities(words) also appear in its corresponding article body, this comment itself, its corresponding article and the selected entity are re-composed as an article-comment-answer sample. After the above data preprocess, finally, 143, 906 article-comment-answer samples are retained in NYT dataset.

2) HUMAN ASSOCIATION DATASET

As a counterpart of machine association dataset, a large scale dataset of human association named SWOW-EN [2] is adopted. Specially, each cue word in SWOW-EN is accompanied with three different response words referred as the first (R1), second (R2) and third (R3) response respectively. Among them, R2 and R3, the second and third responses, are considered to be weaker or remoter responses. The dataset SWOW-R1 counting only R1 response includes

[1] The NYT comments are available at https://www.kaggle.com/aashita/nyt-comments/version/13 and the corresponding articles are collected from the URLs the website has provided. The data are collected from January 2017 to May 2017 and from January 2018 to April 2018.

TABLE 1. Statistics of Machine association dataset, Human association dataset and human-judged Similarity datasets. In the first line, # of samples represents the number of passage-query-answer samples.

Dataset	Name	# of Samples(Machine) # of Words(Human/Similarity)
Machine	CNN	387,420
	NYT	143,906
Human	SWOW-R1	12176
	SWOW-R123	12217
Similarity	SimLex-999	999
	WordSim-353 similarity	203
	MTURK-771	771
	WordSim-353 relatedness	252
	Radinsky2011	287
	MEN	3000
	RG1965	65
	Silberer2014	7576

totally 12, 176 words, and the dataset SWOW-R123 counting all the three responses consists of 12, 217 words respectively, which are listed in Table 1. To make it as objective as possible, the cue-response similarity in human association dataset is also computed through the approaches proposed in IV for machine word association.

3) SEMANTIC SIMILARITY DATASETS

As discussed before, the semantic similarity is critical for the investigation of machine association behavior. Besides comparing to human association dataset, the cue-response similarity derived from the machine association network is also evaluated by using 8 standard human-judged similarity datasets, which follows the way in [2]. Different from the human association dataset, whose cue-response similarity is derived from AS, PPMI and RWS measures, the similarities in the 8 datasets are collected from human participant, so they are desirable to be regarded as ground truth for comparison with those of the machine word association. The 8 human-judged similarity datasets include: SimLex-999 [31], WordSim-353 similarity [32], the MTURK-771 data [33], the WordSim-353 relatedness dataset [32], the Radinsky2011 data [34], the MEN data [35], the popular RG1965 dataset [36], and the Silberer2014 dataset [37]. The details of word number in the 8 similarity datasets are also shown in Table 1.

B. RESULTS OF ASSOCIATION DETECTION

The association words can be obtained once the framework is trained on the machine association datasets. The accuracy performances of predicting the answer correctly on the CNN and NYT datasets with different attention algorithms are shown in Table 2. To further assess the effectiveness of the model, the accuracy performances on CNN dataset of *Attentive Reader* adopted by framework are compared to various reading comprehension algorithms, as shown in Table 3, where the hyperparameter settings for algorithm training follow the work [38]. It can be seen that among all the algorithms, the *Attentive Reader* has acceptable performance. Specifically, the simplest *dot* attention mechanism gives relatively high accuracy, which keeps in line with the conclusion in [24]. Considering that the main purpose of the algorithm is to detect word association rather than improving the answering performance, therefore, the relatively moderate accuracy of the 3 alternatives of *Attentive Reader* algorithm are viewed enough to effectively obtain informative association from the learned attention weights. Similar claim can be stated on NYT dataset.

TABLE 2. Model Accuracy (%) on CNN and NYT datasets.

	CNN		NYT	
Model	Train	Test	Train	Test
Attentive Reader-Bilinear Attention	81.7	72.0	74.5	70.6
Attentive Reader-MLP Attention	74.7	68.4	70.7	66.3
Attentive Reader-Dot Attention	83.3	72.3	76.5	68.1

TABLE 3. Accuracy Comparison of RC algorithms on the CNN dataset. Results marked † are from [38].

Model	Train	Test
Deep LSTM Reader†	-	57.0
Impatient Reader†	-	63.8
MemNets†	-	66.8
AS Reader†	-	69.5
Attentive Reader ([21])†	-	72.4
DER Network†	-	72.9
Iterative Attentive Reader†	-	73.3
Stanford AR (relabeling)†	-	73.6
EpiReader†	-	74.0
AoA Reader†	-	74.4
ReasoNet†	-	74.7
BiDAF†	-	76.9
GA Reader†	-	77.9
Attentive Reader-Bilinear Attention	81.7	72.0
Attentive Reader-MLP Attention	74.7	68.4
Attentive Reader-Dot Attention	83.3	72.3

Next, the machine association network is constructed according to method of section III. To be clear, the machine association networks extracted from CNN dataset are denoted as CNN-Bilinear, CNN-MLP and CNN-Dot, respectively according to the 3 attention algorithms described in section III-B. So is done for machine association network on NYT dataset. The number of words (nodes) in the resulting machine association networks are approximate more than 10k, which is shown in Table 4.

C. EVALUATION METRICS

In an attempt to figure out whether the semantic property of machine association network resembles that of a human association network, the *Pearson correlation coefficient* between semantic similarity of machine association and human association is utilized as measuring metric. Taking the numerical value of similarity between arbitrary pair of nodes in the association network as a sample drawn from a probability distribution, each association network would be characterized by a random variable. That is to say, all the semantic similarities of association network would be regarded from a specific distribution. For two similarity random variables S_i and S_j of association networks G_i and G_j, the correlation coefficient r_{ij}

TABLE 4. Number of overlap words between Association Networks.

Dataset	Number of Words	CNN-Blinear	CNN-MLP	CNN-Dot	NYT-Bilinear
		10653	10384	10561	10039
		overlap	overlap	overlap	overlap
SWOW-R1	12176	4938	4920	4904	4801
SWOW-R123	12217	4938	4920	4904	4825
MEN	3000	1503	1485	1487	1392
MTURK-771	771	351	350	350	336
Radinsky2011	287	54	54	54	61
RG1965	65	38	38	38	34
Silberer2014	7576	2388	2388	2372	1761
SimLex-999	999	509	508	508	495
WordSim-353 relatedness	252	83	83	83	81
WordSim-353 similarity	203	98	98	98	90

is defined as

$$r_{ij} = \frac{Cov(S_i, S_j)}{\sigma_i \sigma_j} \tag{18}$$

In general, the higher the coefficient r_{ij} between machine and human association networks, the closer the word similarity distribution between them [2], and the more value of the automatically extracted association words.

Note that in order to measure the Pearson correlation coefficients appropriately, only words that are present both in machine association network and human judged similarity datasets, or machine association network and human association network are involved. The words only appear in one network but not the other one are ignored. The number of words shared by 4 constructed machine association networks and 2 human association networks, as well as 8 standard similarity datasets, are shown in Table 4.

D. RESULTS OF CONSISTENCE ESTIMATION

First of all interests, the Pearson correlation coefficients between two human association networks (SWOW-R1 and SWOW-R123) and 3 machine association networks (CNN-Bilinear, CNN-MLP and CNN-Dot) are shown in Fig. 2. As can be seen, the correlation coefficients between human and machine association networks increase with similarity measurements from AS to PPMI and then RWS. Recall that the similarity measurement AS considers only the intermediate neighbours while the RWS consider multi-hop subgraph around a word node, it can be concluded that the human and machine association networks are more similar to each other when more association words are involved in the similarity estimation. Here, the low correlation coefficients using AS measurement also agree with Deese's point [39] that the simple frequency of response word is not an ideal measure of semantic similarity. Meanwhile, the correlation coefficients are observed independent with special attention mechanism since they are much similar across 3 different attention algorithms. This result is important for the task of automatic association word detection in that the association words extracted from text datasets are mainly determined by their association relation and not biased with any special detection algorithms.

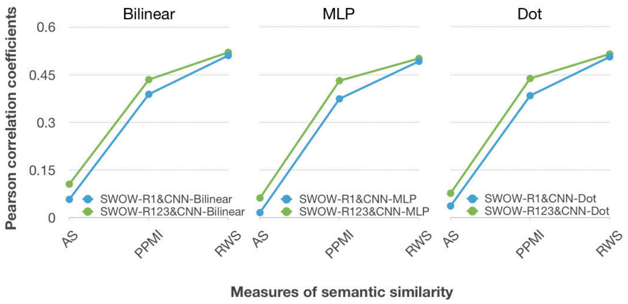

FIGURE 2. Pearson correlation coefficients between similarities of machine association networks and human association networks.

With the above results, the RWS similarity measurement is mainly selected to estimated the association consistence between human and machine association networks. In particular, the average correlation coefficients with RWS across 3 attention mechanisms is as high as 0.508. That is, *the human and machine association networks are significantly consistent to some extent.*

Another observation in Fig. 2 is that the human association network SWOW-R123 has higher correlation coefficients than that of SWOW-R1. The result is not surprising, since, taking the machine association network CNN-Bilinear as example, it is closer to SWOW-R123 in that each word node in CNN-Bilinear has as many as 30 association words as it neighbours, while the SWOW-R123 has 3 association neighbours and SWOW-R1 has only 1 association neighbour. Thus, the CNN-Bilinear is more associative consistent with SWOW-R123 than SWOW-R1.

To further investigate the properties of the obtained machine association networks, similar experiment is conducted between machine association networks and 8 human-judged similarity datasets. The results of Pearson correlation coefficients on CNN dataset are shown in Fig. 3. The correlation coefficients using RWS measurement range in $0.22 \sim 0.56$. In general, the overall correlation results are similar to those between machine and human association networks already discussed in Fig. 2, regardless of which human judgement dataset or which attention mechanism is employed. This indicates the semantic similarity of the machine association network is relatively stable with respect

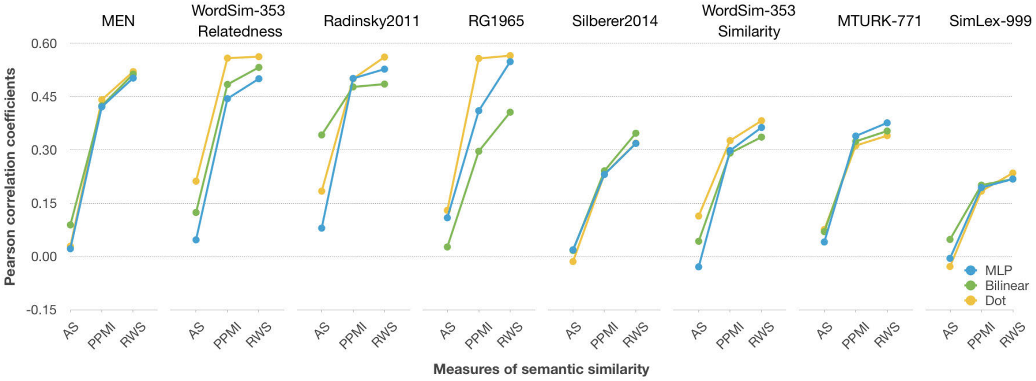

FIGURE 3. Pearson correlation coefficients between three machine association networkson CNN datasets with different attention mechanism and eight human-judged similarity datasets.

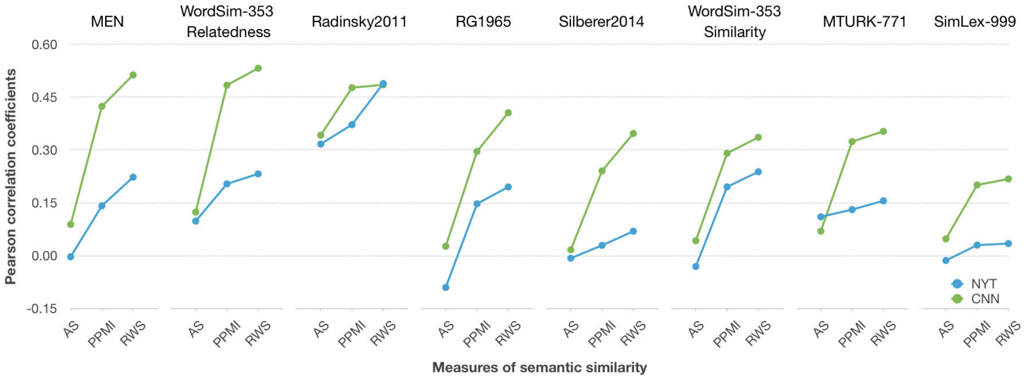

FIGURE 4. Pearson correlation coefficients between human-judged similarity datasets and machine association networks on CNN and NYT datasets.

to human judgement, which further concretes the conclusion that *machine association network is generally consistent with human common sense*. This finding would helpful to the practical usage of machine association network.

A notable question in Fig. 3 is why the coefficient for SimLex-999 is lower than other human-judgement similarity datasets. The reason for this fact lies in the nature of SimLex-999 dataset, in which participants are asked to only judge similarities of word pairs and ignore their relatedness, whereas most other human judgement datasets record relatedness of word pairs. For one thing, in human semantic cognition, the role of similarity is more an empirical result, the ratings of which have been realized less reliable than relatedness ratings [31]. For another, the same pattern of low coefficients in SimLex-999 has been observed in human association network [11]. That is, the similarities captured by human association network also perform the worst when compared with the SimLex-999 dataset. Moreover, the result provides an insight that machine association network tends to construct a structure of lexicon relying more on relatedness than similarity, which keeps in line with the conclusion from a fMRI study [40]: the structure of human mental lexicon is mapped in a more thematic rather than taxonomic way based on similarity.

However, the coefficients for *WordSim-353 Similarity*, another human judged dataset claiming to record only similarity rather than relatedness, are relatively high. It is largely attributed to the fact that the original *WordSim-353* dataset is split into related and similar items by post-hoc raters. As such, the *WordSim-353 Similarity* dataset might not consist of "pure" similarity judgement and hence the misclassified related word pairs may be the result of the relatively high correlation.

E. ASSOCIATION IN DIFFERENT TEXT

Another interesting question is whether the specificity of text dataset upon which the machine association network is constructed has effect to the association property. To answer this question, the correlation coefficients between machine association networks from two text datasets, CNN-Blinear and NYT-Blinear, and human-judged similarity datasets are calculated and plotted in Fig. 4. The CNN-Blinear and NYT-Blinear are regarded comparable in that they have similar word overlap with human-judgement similarity datasets, as well as using the same *Bilinear* attention algorithm.

It can be seen that the correlation of CNN dataset is always higher than that of NYT dataset, no matter which similarity measurement is employed. It is believed that the inherent logical association between article and comments of the two datasets accounts for this difference. Recall that the CNN

TABLE 5. Example cue-response word pairs in NYT-Bilinear association network.

cue	response	association strength (10^{-2})
Trump	president	1.915
	republican	1.028
	America	0.728
	Russia	0.703
	White House	0.689
China	Trump	0.847
	Chinese	0.506
	America	0.309
	president	0.256
	nuclear	0.232
Putin	Russia	0.080
	president	0.046
	Italy	0.039
	White House	0.032
	Trump	0.023

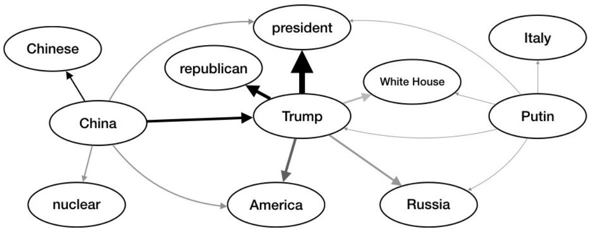

FIGURE 5. Example machine association network constructed according to word pairs in Table 5. The thickness of the directed edge from cue word to response word indicates their association strength.

dataset consists of pairs of news article and its brief summary, that means the information in summary is usually covered by the article. Therefore, as the role of comment for RC model training, the summary of CNN dataset is directly and closely related to the article. The association between them is easy to be captured by attention mechanism through assigning significant attention weight. In contrast, the NYT dataset contains comments from thoughts of various audiences, which may or may not directly refer the information contained in article. Hence, the inherent logical association between comments and articles are obviously different in CNN and NYT datasets, which leads to the superior correlation coefficient on CNN data. This implies a new interesting discussion of strong and remote association, which would be explored as further work.

F. CASE STUDY

In addition to the analysis of association consistence between machine and human association networks, to make it more explicative, Table 5 shows some cue-response word examples extracted from NYT dataset, where the cue words and response words are listed in the first two columns, and the association strengths between them, namely f_{ij} defined in (13), given in the last column. A toy machine association network constructed based on Table 5 is also plotted in Fig. 5.

In general, the association relations between cue-response words in Fig. 5 are quite in line with human intuition. For example, the cue word *Trump* has 5 neighbouring response words, which coincides the well-known fact that *Donald Trump* is a US *Republican* politician and the 54th *president* of *America*, and now lives in the *White House*. Meanwhile, *Trump* is also associated with *Russia,* apparently because of Russia's political ties to the United States. It is noticeable that the association strength from *Trump* to *president* is larger than that from *China* to *president*, or from *Putin* to *president*. It is because that NYT, The New York Times, is daily published in New York of the United States, in which the articles and comments of U.S. President *Trump* are more frequently mentioned than those of *China* and *Putin*.

Another point worth mentioning is that, it seems counterintuitive to see the association strength from *Trump* to *Russia* is strong than that from *Putin* to *Russia*. The nature of NYT that it is an American daily accounts for this observation again. There are more articles focusing on the issues between American president, namely *Trump*, and *Russia* than those between other country's president (*Putin* in this case) and *Russia*. The same logic can be applied to explain the case of *China-Trump* and *China-Chinese*, where the association strength between the former is higher than that between the latter, since the topics between *China* and *Trump* are talked about more frequently than those between *China* and *Chinese*.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a neural network based framework for automatic association detection, to feasibly collect association words from plain text. The reading comprehension algorithm and attention mechanism are employed to fulfill the task. The semantic consistence between the machine and human association networks is experimentally verified. This finding provides an insight of understanding the associative property of the extracted association words, and evaluating their potential usage in various research domains.

This work takes just a small step forward to extending the traditional psychological research by incorporating approaches of NLP. The future work will focus on the association detection methodology with word-to-word attention to explore fine-grain association among words. Also, exploring the phenomenon of strong and remote association in machine word association network is another direction of future work.

REFERENCES

[1] S. B. Klein, *Learning: Principles and Applications*. Newbury Park, CA, USA: Sage,, 2018.
[2] S. De Deyne, D. J. Navarro, A. Perfors, M. Brysbaert, and G. Storms, "The 'small world of words' English word association norms for over 12,000 cue words," *Behav. Res. Methods*, vol. 51, no. 3, pp. 1–20, 2018.
[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[4] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019.
[6] Y. Bao, S. Chang, M. Yu, and R. Barzilay, "Deriving machine attention from human rationales," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1903–1913.

[7] N. Pappas and A. Popescu-Belis, "Human versus machine attention in document classification: A dataset with crowdsourced annotations," in *Proc. 4th Int. Workshop Natural Lang. Process. Social Media*, 2016, pp. 94–100.

[8] S. Mednick, "The associative basis of the creative process," *Psychol. Rev.*, vol. 69, no. 3, p. 220, 1962.

[9] S. A. Mednick, "The remote associates test," *J. Creative Behav.*, vol. 2, no. 3, pp. 213–214, 1968.

[10] J. J. Jenkins, "The 1952 minnesota word association norms," in *Norms of Word Association*. Amsterdam, The Netherlands: Elsevier, 1970, pp. 1–38.

[11] S. De Deyne, A. Perfors, and D. J. Navarro, "Predicting human similarity judgments with distributional models: The value of word associations," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2016, pp. 1861–1870.

[12] B. Schloss and P. Li, "Disentangling narrow and coarse semantic networks in the brain: The role of computational models of word meaning," *Behav. Res. Methods*, vol. 49, no. 5, pp. 1582–1596, 2017.

[13] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," *IJcAI*, vol. 7, pp. 1606–1611, Jun. 2007.

[14] S. Hassan, "Measuring semantic relatedness using salient encyclopedic concepts," Univ. North Texas, Denton, TX, USA, Tech. Rep., 2011.

[15] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans. Knowl. Discovery Data*, vol. 2, no. 2, p. 10, 2008.

[16] T. Hughes and D. Ramage, "Lexical semantic relatedness with random graph walks," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, 2007.

[17] R. Ghaeini, X. Z. Fern, and P. Tadepalli, "Interpreting recurrent and attention-based neural models: A case study on natural language inference," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018.

[18] D. Alvarez-Melis and T. S. Jaakkola, "A causal framework for explaining the predictions of black-box sequence-to-sequence models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017.

[19] R. Ghaeini, S. A. Hasan, V. Datla, J. Liu, K. Lee, A. Qadir, Y. Ling, A. Prakash, X. Z. Fern, and O. Farri, "Dr-bilstm: Dependent reading bidirectional LSTM for natural language inference," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018.

[20] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1693–1701.

[21] D. Chen, J. Bolton, and C. D. Manning, "A thorough examination of the CNN/daily mail reading comprehension task," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 2358–2367.

[22] H. Zhu, F. Wei, B. Qin, and T. Liu, "Hierarchical attention flow for multiple-choice reading comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[23] C. Clark and M. Gardner, "Simple and effective multi-paragraph reading comprehension," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018.

[24] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[26] S. De Deyne, D. J. Navarro, A. Perfors, and G. Storms, "Structure at every scale: A semantic network account of the similarities between unrelated concepts," *J. Experim. Psychol., Gen.*, vol. 145, no. 9, p. 1228, 2016.

[27] K. A. Hutchison, D. A. Balota, J. H. Neely, M. J. Cortese, E. R. Cohen-Shikora, C.-S. Tse, M. J. Yap, J. J. Bengson, D. Niemeyer, and E. Buchanan, "The semantic priming project," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1099–1114, 2013.

[28] G. Recchia and M. N. Jones, "More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis," *Behav. Res. Methods*, vol. 41, no. 3, pp. 647–656, 2009.

[29] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[30] M. Newman, *Networks: An Introduction*. Oxford, U.K.: Oxford Univ. Press, 2010.

[31] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating semantic models with (genuine) similarity estimation," *Comput. Linguistics*, vol. 41, no. 4, pp. 665–695, 2014.

[32] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *Proc. Hum. Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2009, pp. 19–27.

[33] G. Halawi, G. Dror, E. Gabrilovich, and Y. Koren, "Large-scale learning of word relatedness with constraints," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1406–1414.

[34] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: Computing word relatedness using temporal semantic analysis," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 337–346.

[35] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, "Distributional semantics in technicolor," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2012, pp. 136–145.

[36] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[37] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014, pp. 721–732.

[38] B. Dhingra, H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Gated-attention readers for text comprehension," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1832–1846.

[39] J. Deese, *The Structure of Associations in Language and Thought*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1966.

[40] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, p. 453, 2016.

ZHENG HU received the B.S. degree from the Nanjing University of Posts and Telecommunications, in 2002, and the Ph.D. degree in circuits and systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2008. He is currently working with the State Key Laboratory of Networking and Switching Technology, BUPT. His current research interests include cyber space user behavior modeling, analyzing, cognition, and intelligent application systems.

JIAO LUO received the B.S. degree in photoelectric information science and engineering from the Beijing University of Posts and Telecommunications, in 2018, where she is currently pursuing the M.S. degree in information and communication engineering. Her research interests include word association and natural language generation.

CHUNHONG ZHANG received the B.Eng. degree in telecommunication engineering, in 1993, the M.Eng. degree in information technology, in 1996, and the Ph.D. degree in computer science, in 2013. She was a Visiting Scholar with the Illinois Institute of Technology, in 2015. She is currently a Lecturer with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. Her research interests include data mining, recommendation, nature language processing, and computer vision. She has served as a TPC Member for international conferences on BigCom.

WEI LI received the Ph.D. degree in electrical engineering from the University of Victoria, Canada, in 2004. He is currently an Assistant Professor and the Director of the Laboratory for Intelligent Networks and Systems, Northern Illinois University, DeKalb, IL, USA. His research interests include intelligent networks, the Internet of Things, smart grid, computer vision and cognition, machine learning, and artificial intelligence.